ARTICLE

Leveraging Multi-ethnic Evidence for Mapping Complex Traits in Minority Populations: An Empirical Bayes Approach

Marc A. Coram,[1,8,9] Sophie I. Candille,[2,8] Qing Duan,[3] Kei Hang K. Chan,[4] Yun Li,[3,5] Charles Kooperberg,[6] Alex P. Reiner,[6,7] and Hua Tang[2,*]

Elucidating the genetic basis of complex traits and diseases in non-European populations is particularly challenging because US minority populations have been under-represented in genetic association studies. We developed an empirical Bayes approach named XPEB (cross-population empirical Bayes), designed to improve the power for mapping complex-trait-associated loci in a minority population by exploiting information from genome-wide association studies (GWASs) from another ethnic population. Taking as input summary statistics from two GWASs—a target GWAS from an ethnic minority population of primary interest and an auxiliary base GWAS (such as a larger GWAS in Europeans)—our XPEB approach reprioritizes SNPs in the target population to compute local false-discovery rates. We demonstrated, through simulations, that whenever the base GWAS harbors relevant information, XPEB gains efficiency. Moreover, XPEB has the ability to discard irrelevant auxiliary information, providing a safeguard against inflated false-discovery rates due to genetic heterogeneity between populations. Applied to a blood-lipids study in African Americans, XPEB more than quadrupled the discoveries from the conventional approach, which used a target GWAS alone, bringing the number of significant loci from 14 to 65. Thus, XPEB offers a flexible framework for mapping complex traits in minority populations.

Introduction

Genome-wide association studies (GWASs) combined with sequencing have fundamentally transformed the field of human complex-trait genetics; since 2007, thousands of loci have been identified for a broad spectrum of traits and diseases.[1] However, the success of GWASs has been primarily confined to populations of European descent. Minority individuals, such as African-American (AA) and Hispanic-American (HA) individuals, who together represent 28% of the US population (US 2010 census), are prominently missing from many studies. For example, the largest GWAS of plasma lipids (the Global Lipids Genetics Consortium or GLGC) included over 188,000 individuals of European ancestry and identified more than 150 loci associated with lipid traits.[2] By comparison, the largest published discovery GWAS for lipid traits in US minorities comprised only ~8,000 AA and ~3,500 HA women.[3] In part for this reason, minority samples have predominantly been utilized for replication, generalizability, and/or fine mapping rather than primary GWAS discovery.[4–9] Although several large-scale GWASs dedicated exclusively to AA and HA populations are underway,[10] these studies are nonetheless underpowered on the basis of the empirical evidence that complex traits and diseases are often influenced by a myriad of variants, each with moderate genetic effects requiring a large sample size to be uncovered.[11] The sample-size limitation and analytic challenges are likely to be exacerbated among minority populations as we move toward sequencing-based studies. As a result, there is a widening gap in our knowledge about genetic risk factors for complex diseases across racial and ethnic groups.

Much of the success of GWASs in European populations owes to the ability to combine cohort-specific summary results from a large number of studies via meta-analysis techniques. It is therefore reasonable to hypothesize that the efficiency of complex-trait association studies in minority populations might be improved when smaller minority samples are analyzed in conjunction with the much larger European and European-American cohorts. Indeed, there is accumulating evidence that, for a variety of complex traits and diseases, there is substantial overlap in trait-associated loci between ethnicities.[7,9,12] For plasma lipid concentration, we previously demonstrated that trait-influencing loci show excess overlap among AA, HA, and European-descent (EU) populations and that loci identified in EU populations explain a disproportionate amount of the phenotypic variance in both AA and HA populations.[3] Despite such overlap, conventional meta-analysis approaches—both fixed-effects and random-effects models—are not appropriate for combining data across race and ethnicity. These approaches assume that the underlying disease variants and allelic effects are identical

[1]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA; [2]Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA; [3]Departments of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA; [4]Laboratory of Molecular Epidemiology and Nutrition, Department of Epidemiology, Brown University, Providence, RI 02912, USA; [5]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA; [6]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [7]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA
[8]These authors contributed equally to this work
[9]Present address: Google Inc., Mountain View, CA 94043, USA
*Correspondence: huatang@stanford.edu
http://dx.doi.org/10.1016/j.ajhg.2015.03.008.

or similar, but heterogeneity in genetic architecture between ethnicities is well documented. A well-known example is the APOE ε4 allele, which confers considerably higher risk of Alzheimer disease in Japanese than in AA individuals.[13] Moreover, because European sample sizes are often much larger than those of minority individuals, the meta-analysis framework (which weighs individual studies according to sample size or the inverse of the estimated variance) preferentially identifies loci showing association in Europeans while sacrificing power to detect minority-specific risk loci.

On the basis of these considerations, we propose an empirical Bayes (EB) approach[14,15] designed to elucidate the genetic architecture of complex traits in a minority ethnic population while adaptively incorporating GWAS information from other ethnicities. We reason that the general relevance of GWAS results across ethnicities is often unknown a priori and might depend on both the genetic architecture of a specific trait and the evolutionary relationship between populations; however, it can be gauged empirically on the basis of the genome-wide consistency in association evidence. In other words, if the underlying genetic basis of a trait is similar between two ethnicities, and the genetic architecture is polygenic, we would observe greater overlap in loci showing trait association in the two populations than expected by chance.

We show through simulations that our proposed cross-population empirical Bayes (XPEB) approach behaves sensibly. When the underlying trait-associated loci largely coincide, XPEB effectively combines the two populations and approximates the power of a fixed-effects meta-analysis; at the other extreme, when the genetic bases are entirely population specific, XPEB only uses the population of interest (referred to as the target population). When genetic architecture partially overlaps, XPEB outperforms fixed-effects and random-effects meta-analyses, as well as the approach of considering only the target population.

Material and Methods

Data and Model

Consider a situation in which GWASs have been conducted for a trait in a *target* (e.g., AA) population and a *base* (e.g., EU) population and in which SNP-level summary statistics are available in both target and base populations at M markers. The goal is to identify risk variants in the target population. We make no assumption with regard to the genetic similarity between the target and base populations but require that the target GWAS and base GWAS be performed on non-overlapping individuals.

Let S_m and S'_m be the test statistics (to be specified) of marker m in the target and base populations, respectively. Within each population, the distribution of the test statistic is modeled as a mixture with $J+1$ components, of which the first component represents the null distribution of the test statistics at markers not associated with the trait, and the remaining components represent the alternative distributions of the test statistics at trait-associated markers. We introduce an unobserved Bernoulli random variable, D_m, at each marker m, such that $D_m = 1$ with a probability of π_1 and indicates that the test statistic observed in the target population is drawn from a non-null component. Thus, in the target population, the test statistics have a density of

$$g(s) = (1-\pi_1)g_0(s) + \pi_1 \sum_{j=1}^{J} p_j g_j(s), \qquad \text{(Equation 1)}$$

where π_1 represents the proportion of non-null markers and $\boldsymbol{p} = (p_1, \cdots p_J)$, $(\sum_{j=1}^{J} p_j = 1)$ represents the relative proportions of non-null SNPs that are associated with the phenotype at varying degrees; $g_j(s)$ $(j = 1,\ldots,J)$ are densities corresponding to the non-null components, which are described briefly in the next section and in detail in Appendix A. For notational simplicity, we denote the weighted sums of the non-null components in Equation 1 as $\tilde{g}(s)$. To formulate the corresponding density in the base population, we assume that the distributions of allelic effects are similar between populations. In other words, on average, the relative fraction of loci with strong and weak effects—measured by the proportion of phenotypic variance explained by each locus—is similar between populations, even though any given locus is not necessarily shared. Therefore, we let $\boldsymbol{p}$ be shared between target and base populations and modify $g_j(s)$ to account for the differential target and base GWAS sample sizes (Appendix A). We allow the overall contribution of the non-null component, π_1 and π'_1, to differ between the target and base populations because the number of "detectable" loci in a GWAS depends on the sample size. It follows that the density of S'_m differs from Equation 1 in the null proportion, π'_1, and the shape of the non-null components, $g'_j(s)$:

$$g'(s) = (1-\pi'_1)g_0(s) + \pi'_1 \sum_{j=1}^{J} p_j g'_j(s). \qquad \text{(Equation 2)}$$

Additionally, let D'_m denote the unobserved random variable that determines whether the test statistic in the base population is drawn from the null or the non-null component.

The key insight underlying the proposed approach is that, at each marker m, the random variables (D_m and D'_m) are not independent if the genetic architecture overlaps between the target and base populations, and the degree of overlap can be estimated on the basis of the empirical *genome-wide* joint distribution of the observed GWAS statistics (S_m and S'_m) across all markers. Specifically, we model the conditional probabilities, $P(D_m = 1|D'_m = 0)$ and $P(D_m = 1|D'_m = 1)$, as constant across the genome and denote these two parameters as κ_0 and κ_1, respectively. Assuming that S_m and S'_m are independent given D_m and D'_m, respectively, the likelihood of observing a test statistic s_m is

$$\begin{aligned} g^*(s_m) &= \sum_{a\in\{0,1\}}\sum_{b\in\{0,1\}} P(D'_m = a \,|\, S'_m) P(D_m = b \,|\, D'_m = a) P(S_m \,|\, D_m = b) \\ &= (1-\nu'_m)[(1-\kappa_0)g_0(s_m) + \kappa_0 \tilde{g}(s_m)] \\ &\quad + \nu'_m[(1-\kappa_1)g_0(s_m) + \kappa_1 \tilde{g}(s_m)], \end{aligned} \qquad \text{(Equation 3)}$$

where $\nu'_m = P(D'_m = 1|S'_m)$, and the overall likelihood of S (under linkage equilibrium), is simply

$$\text{lik}(S) = \prod_m g^*(s_m). \qquad \text{(Equation 4)}$$

In Equation 3, $1-\nu'_m$ has the interpretation of the local false-discovery rate (locfdr) for testing the statistical association between a SNP and the trait in the base population.[16] Maximizing the likelihood of Equation 4 provides estimates for the model parameters in

Equation 3; these parameters are used for computing the posterior probability, $P(D_m = 0|S_m, S'_m)$, which will be used as the cross-population locfdr of genotype-trait association in the target population.

Implementation of XPEB

The model described above is applicable to any GWAS summary statistics. Two natural choices are t statistics and p values, which differ only in that t statistics, but not p values, inform the direction of association. However, t statistics are not always available, and even when they are, merging two lists of t statistics can be prone to error because of the ambiguity in the designation of the reference alleles. In contrast, p values are the most widely shared summary statistics and do not rely on consistent coding of the reference alleles; therefore, we choose to base our implementation on p values. More precisely, in the likelihood of Equation 3, we use χ^2_1 statistics for S and S', which can be derived from either p values or t statistics through a quantile transformation. It is then reasonable to use a χ^2_1 distribution as the null component (g_0) in the mixture densities of Equations 1 and 2 and to use a mixture of non-central χ^2_1 distributions to represent the non-null components. We introduce non-null basis functions $g_j(s)$ (with $j = 1,\ldots,J$ and $J = 25$), each of which is itself a continuous mixture of non-central χ^2_1 distributions. These basis functions are fixed a priori with masses concentrating at different values (Figure S1), reflecting the expectation that allelic effects under a polygenic architecture might span a broad spectrum, as measured by the proportions of phenotypic variance explained.[11] Each basis function in the base and target populations is scaled to reflect the unequal sample sizes between the target and base GWASs. Details on the formulation of these basis functions are described in Appendix A. We have experimented with using $J = 50$ in simulations and found negligible differences in performance. To estimate the model parameters, we developed an EB approach that proceeds in two stages, which are described next. The first stage deconvolves the mixture density in Equations 1 and 2 by estimating the mixture proportions, and the second stage estimates parameters κ_0 and κ_1 by maximizing the likelihood, lik(S), of Equation 4.

The first stage estimates the mixture proportions (π_1, π'_1, and $\boldsymbol{p}$) in Equations 1 and 2 by assuming that the trait-influencing loci are independently distributed within the target and base populations. One way to estimate these parameters is to maximize the marginal likelihood via the EM algorithm.[17] However, theoretical studies[18] and our own simulations have found that mixture proportions estimated this way can be undesirably sensitive when the non-null probabilities in target or base populations are close to 0; such a situation can arise either because the trait is affected by just a few loci or because the signal in the GWAS is weak as a result of small sample sizes. As an alternative, we implemented a hierarchical Bayesian model, implemented via a Markov Chain Monte Carlo (MCMC) algorithm, to explore multiple plausible deconvolutions (see Appendix A). We use the posterior mean as the estimated mixture proportions. Once the mixture proportions are estimated, we can compute ν'_m for each marker by substituting the estimated parameters as follows:

$$\nu'_m = P(D'_m = 1 \mid S'_m) = \frac{\widehat{\pi}'_1 \sum_{j=1}^{J} \widehat{p}_j g'_j(s)}{(1-\widehat{\pi}'_1) g_0(s) + \widehat{\pi}'_1 \sum_{j=1}^{J} \widehat{p}_j g'_j(s)}. \quad \text{(Equation 5)}$$

Analogously, the target-only posterior probability of association is defined as $\nu_m = P(D_m = 1|S_m)$. The goal of the second stage of the algorithm is to adaptively integrate information from the target and base populations. We use an EB approach to estimate parameters κ_0 and κ_1, which maximizes the likelihood function of Equation 4 while being subject to the constraints $0 \le \kappa_0 \le \kappa_1 \le \varsigma < 1$. In the implementation, we set $\varsigma = 0.9$ and express a belief that the genetic architectures in two populations are seldom identical; this restriction improves numerical stability without substantial loss of efficiency. Finally, the XPEB locfdr for the SNP association in the target population, ω_m, is computed as

$$\omega_m = P(D_m = 0 \mid S_m;\ S'_m) = \frac{\left[(1-\upsilon'_m)(1-\widehat{\kappa}_0) + \upsilon'_m(1-\widehat{\kappa}_1)\right] g_0(s_m)}{\widehat{g}^*(s_m)}, \quad \text{(Equation 6)}$$

where $\widehat{g}^*(s_m)$ is the likelihood of observing a test statistic s_m, defined in Equation 3 and evaluated by the substitution of estimated parameter values. Note that the cross-population posterior estimates reduce to a one-population model in the special case that $\kappa_0 = \kappa_1$; in other words, when XPEB does not detect evidence of overlapping genetic architecture, results in the base GWAS are ignored.

The model and implementation described thus far assume that markers are independent; that is, there is no linkage disequilibrium (LD) between SNPs. LD affects two aspects of the procedure. First, during estimation of the mixture proportions, the presence of LD can lead to an over-estimation of non-null proportions, π_1 and π'_1. To overcome this problem, we introduce an LD-trimming step and estimate these parameters in two iterations (see Appendix A). A second effect of LD is that differential LD patterns between the target and base populations can lead to an under-estimate of κ_1, as we explain in the Results and demonstrate through simulations. However, such a downward bias in κ_1 leads to slightly conservative estimates of locfdr but does not inflate false-positive estimates. Hence, the uncorrected $\widehat{\kappa}_1$, estimated on full data, is used in all subsequent simulation and data analyses.

Simulations

We used two sets of simulation experiments to evaluate the performance of the proposed approach. The first set of experiments directly simulated test statistics on the basis of theoretically computed non-centrality parameters; this approach did not simulate individual genotypes and assumed that all markers were independent (i.e., no LD). The simulation proceeded in three stages. The first stage modeled the genetic architectures of the trait, which is assumed to have an additive genetic architecture specified by the total heritability (h^2) and the number of causative variants in each population (K), of which a fraction of δ loci are shared between the target and base populations. We emphasize that although both parameters δ and κ_1 measure the degree of overlap, δ is a simulation parameter and is allowed to range between K/M (independence) and 1 (complete overlap), whereas κ_1 is a model parameter with the constraints specified by XPEB. Under such a trait model, stage two simulated the allelic effect and additive variance attributable to each causal locus. At target-specific or base-specific trait-associated loci, the allelic effects, λ, were sampled independently from a standard normal distribution. At a shared trait-associated locus, the allelic effects in target and base populations were sampled from a bivariate normal distribution, $N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & r\\ r & 1\end{bmatrix}\right)$, where $r = 0.7$. The phenotypic variance attributable to locus m can be computed as $\xi_m = C\lambda^2_m f_m(1-f_m)$, where f denotes the allele

frequency and C is a normalizing constant such that the total additive variance matches the pre-specified heritability. The allele frequencies at M simulated markers were sampled without replacement from the 1000 Genomes Project variants (phase 1 release v.3).[19] EU populations were used for the base GWAS, and populations of African ancestry were used for the target GWAS; minor allele frequencies (MAFs) greater than 1% were required in both populations. Given ξ_m and under the assumption that ξ_m is small, the GWAS test statistic (on a Z scale) at this causative SNP was sampled from a normal distribution with mean $\xi_m\sqrt{n}$ (where n is the sample size) and variance 1. The remaining $M - K$ SNPs were assumed to not be associated with the trait, and the test statistics were sampled from a standard normal distribution.

The simulated dataset—consisting of a list of Z statistics in each of the target and base populations—was analyzed via four strategies: (1) p values in the target population alone (target only), (2) locfdr estimated by the proposed EB approach combining target and base populations (XPEB), (3) fixed-effects meta-analysis implemented in PLINK (meta-FE),[20] and (4) random-effects meta-analysis implemented in METASOFT (meta-RE, corresponding to the "PVALUE_RE2" in METASOFT output), which has been demonstrated to achieve higher power than the classic random-effect meta-analysis method.[21] For meta-analysis, the SEs for the estimated allelic effects from a linear regression model are given by $\sigma_m^2 = (1 - \xi_m)/[2f_m(1 - f_m)(n - 2)]$. Hence, the estimated allelic effects were computed as $Z_m\sigma_m$. XPEB declares a positive finding by using a locfdr threshold, whereas the other three methods use a p value threshold. Decisions from each of these strategies were compared against the simulation gold standard, defined as the set of risk loci in the target population, and the numbers of false-positive and false-negative discoveries were recorded. We note that the stringency of a p-value-based testing procedure is not directly comparable to that of a procedure based on false-discovery rate (FDR).[22] Furthermore, although a p value of 5×10^{-8} is widely adopted as a significance criterion in GWASs, studies using FDR-based approaches do not always use the same significance levels. With these considerations in mind, we compared the numbers of true loci discovered while adjusting the p value or locfdr thresholds such that each of the four tests had the same number of false-positive discoveries. We then varied the calibration number of false-positive discoveries so that the relative performance of these methods could be evaluated over a range of statistical stringencies. We emphasize, however, that in analyzing real data, the significance criterion should be fixed prior to the analysis.

The approach described above is computationally efficient and allowed us to explore the effects of different aspects of model parameters. However, it is difficult to incorporate a realistic pattern of LD under this framework; therefore, we adopted a complementary approach that directly simulates individual-level genotype data and traits. Specifically, we used YRI (Yoruba in Ibadan, Nigeria) and CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) genotypes from HapMap phase 3 as the prototypes for the target and base populations, respectively.[23] We used HAPGEN2, which introduces mutation and recombination between haplotypes, to simulate ten CEU (base) cohorts and one YRI (target) cohort, each of which included 10,000 individuals.[24] We restricted all simulations and analyses to ~727,348 SNPs, which are on the Illumina Human 1M array and have a MAF greater than 1% in both the simulated target and base populations. Instead of simulating haplotypes that span an entire chromosome, we generated independent windows of 25 linked SNPs (i.e., SNPs within a window inherit the LD pattern in the HapMap sample, whereas SNPs that reside in different windows are unlinked). We emphasize that these 25-SNP windows (spanning a median of 79 kb) captured the essential and realistic features of LD in the population.[25] The motivation for simulating independent windows was to achieve unambiguous definition of false-positive and true-positive findings: a window in which at least one SNP is deemed statistically significant was considered a true-positive finding if the window harbored a trait-associated SNP; otherwise, this window was considered a false positive no matter how many SNPs were falsely rejected within the window. The trait-associated loci and the allelic effects were specified similarly as in the previous set of simulation experiments, but there was an additional constraint of at most one causative SNP per window. The positions of causal SNPs were identical across all ten CEU-derived cohorts, and the allelic effects had a correlation of 0.9 between the cohorts. A fraction of δ of the causal SNPs coincided between the CEU-derived and the YRI-derived populations, and cross-ethnic correlations in allelic effects were 0.7 at the overlapping causal SNPs. The individual genotypes, causative SNPs, and allelic effects were then used as input in the program GCTA (Genome-wide Complex Trait Analysis), which simulates quantitative or binary disease traits with the pre-specified heritability.[26] To reduce computational and data-storage burden, we simulated the genotypes of the 11 cohorts once and used these genotype data to simulate multiple sets of quantitative and case-control traits; a similar strategy was used and discussed in Wu et al.[27] For each trait, a GWAS was performed with PLINK for each cohort separately. The four procedures described in the previous simulations of independent loci—target only, XPEB, meta-FE, and meta-RE—were compared. For XPEB, the ten CEU cohorts were meta-analyzed first (with PLINK's fixed-effects model), and the meta-analysis summary statistics were used as the base GWAS. For the fixed-effects and random-effects meta-analyses, the estimated allelic effects and the SEs for the estimated allelic effects from the 11 individual cohorts were provided to PLINK and METASOFT, allowing for better estimates of inter-study variability.

Analysis of Lipid Traits

To apply our method to real data, we analyzed high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, and triglyceride (TG) levels. For the target GWAS, we used results from the Women's Health Initiative SNP Health Association Resource (WHI-SHARe), which consists of 8,153 post-menopausal AA women. Details of the cohort characteristics, phenotype transformation, and GWAS analyses were reported previously.[3] For the base GWAS, we used summary-level statistics reported by GLGC,[28] a meta-analysis of ~100,000 individuals primarily of European descent. As negative controls, we also used European GWAS results in height[29] and body mass index (BMI)[30] for a base GWAS. Because the sample sizes of the target GWAS and base GWAS were used for constructing the non-null basis, $g_j(s)$ and $g_j'(s)$, respectively, SNPs with excessive missing observations were removed (see footnotes "a" and "e" in Table 2). For SNPs whose summary statistics were available in the target GWAS but missing in the base GWAS, ν_m was computed analogously as in Equation 5, and $1 - \nu_m$ was reported as the locfdr. Regions harboring SNPs with locfdr < 0.05 were considered statistically significant and reported. As validation, we compared the allelic effects at trait-associated loci identified by XPEB to the corresponding allelic effects estimated in an independent cohort of 7,138 AA participants from the NHLBI Candidate-Gene

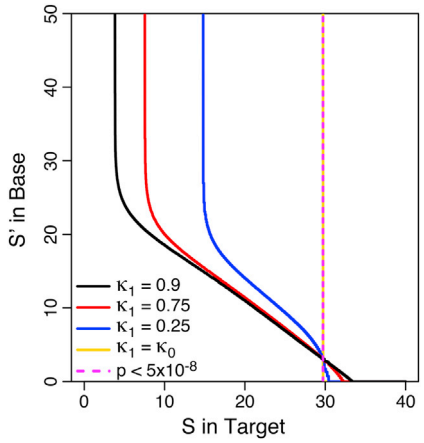

Figure 1. Decision Boundary for XPEB versus That of the Conventional Single-Population Approach

S (x axis) and *S′* (y axis) represent the chi-square test statistics in the target and base GWASs, respectively. The dashed magenta vertical line corresponds to the conventional approach, which uses the target GWAS alone and a genome-wide significance of $p = 5 \times 10^{-8}$. Solid curves delineate the decision boundaries constructed by XPEB, in which the overlap in genetic architecture (κ_1) varies; the yellow line labeled $\kappa_1 = \kappa_0$ corresponds to the case where trait-associated loci are independently distributed in the base and target populations. All other model parameter values are taken from the estimates obtained from the LDL data.

Association Resource (CARe) Study and in a cohort of 3,587 HA participants from WHI-SHARe.

Results

The distinction between XPEB and the conventional single-population approach is qualitatively illustrated in Figure 1. The conventional practice rejects the null hypothesis of no association at a SNP if and only if it exceeds a pre-specified threshold (i.e., $p < 5 \times 10^{-8}$, dashed magenta line in Figure 1) in the target GWAS and does not depend on results in the base population. In contrast, the XPEB approach considers the combined evidence of association in both target and base populations (solid lines in Figure 1). The negative slope of the decision boundaries supports the intuition that the burden of proof in the target population can be reduced by the evidence of association in the base population. As a result, the minimal target evidence for rejecting the null hypothesis can be substantially weaker for XPEB. The slope depends on the simulation parameter δ, which represents the degree of overlap in the genetic architecture underlying the trait in the two populations and is empirically estimated as $\widehat{\kappa}_1$ on the basis of genome-wide test statistics. The greater the overlap, the more influence the base GWAS exerts. When little overlap is detected, the XPEB reduces to a target-only approach with a vertical decision boundary that is independent of the base GWAS (yellow line in Figure 1). Regardless of the estimated overlap, target-specific trait-associated variants can be detected with a sufficiently strong statistic in the target GWAS alone, in which case the test statistic in the base population is effectively ignored, and XPEB coincides with the conventional p value decision. In contrast, a SNP cannot be declared significant on the basis of the strength of evidence in the base population alone as long as $\widehat{\kappa}_1 < 1$, indicated by the vertical gap between the y axis and each decision boundary (Figure 1). This asymmetric feature emphasizes our primary goal of identifying relevant trait variants in the target population and distinguishes XPEB from meta-analysis. Additional Figure 1 features, which are not essential to the understanding of XPEB, are discussed in Appendix A.

Simulations with Independent SNPs

The output of XPEB is an estimated locfdr for each SNP. We recommend rejecting all SNPs with a locfdr falling below a pre-specified threshold (e.g., locfdr < 0.05); alternatively, one can successively reject SNPs with increasing locfdrs until the average locfdr of the rejected set, denoted as the Fdr, reaches a pre-specified threshold (e.g., Fdr < 0.05). Table 1 summarizes the number of true and false discoveries under various simulation settings and significance criteria. As expected, the largest locfdr of the rejected SNP set is a conservative estimate of the realized FDR (rFDR). In contrast, the average locfdr of the rejected set tracks well with the rFDR except for $\delta = 1$, in which case the average locfdr of the rejected set is also conservative in comparison to the rFDR. Additional simulations have verified that the conservatism is due to the constraint of $\widehat{\kappa}_1 \leq 0.9$; the estimated locfdr becomes less conservative when the bound increases.

Next, we compared the performance of XPEB with that of three common approaches: (1) p values in the target population alone (target only), (2) a standard fixed-effects meta-analysis (meta-FE), and (3) a random-effects meta-analysis method combining target and base GWASs (meta-RE).[21] SNPs are ranked according to increasing locfdrs under the XPEB approach and according to increasing p values under the single-population and meta-analysis approaches. Figure 2 compares the number of false-positive rejections that a method makes in order to correctly discover a given number of truly causative (true-positive) variants. Each curve can be thought of as a partial receiver operating characteristic; hence, methods characterized by a greater area under the curve (AUC) are preferred over ones with a lower AUC. As expected, the target-only approach was robust against genetic heterogeneity, whereas the random-effects meta-analysis had the greatest power when the genetic architectures were identical across populations (Figure 2). XPEB combined the strengths of both methods.

In a situation where the trait variants are independently distributed in target and base populations, the target-only and XPEB approaches have the desired behavior of using the target GWAS alone and discarding the information

Table 1. True-Positive and False-Positive Counts at Different locfdr Significance Levels in Simulations of Independent Loci

Statistic	Genetic Architecture	Significance Level: 0.01 TP	0.01 FP	0.01 rFDR	0.05 TP	0.05 FP	0.05 rFDR	0.1 TP	0.1 FP	0.1 rFDR
locfdr[a]	null target[b]	0	0	ND	0	0	ND	0	0.02	ND
	$\delta = 0.001$	66.4	0.04	6.0×10^{-4}	82.9	0.40	4.7×10^{-3}	91.2	0.91	9.9×10^{-3}
	$\delta = 0.25$	81.8	0.06	7.2×10^{-4}	103.4	0.59	5.6×10^{-3}	115.1	1.43	1.2×10^{-2}
	$\delta = 0.5$	110.2	0.13	1.2×10^{-3}	141.7	1.15	8.0×10^{-3}	160.6	2.81	1.7×10^{-2}
	$\delta = 0.75$	152.7	0.27	1.7×10^{-3}	202.9	1.99	9.4×10^{-3}	234.8	5.80	2.3×10^{-2}
	$\delta = 1$	208.1	0.03	1.4×10^{-4}	285.2	0.30	1.1×10^{-3}	340.1	0.83	2.4×10^{-3}
Fdr[c]	null target	0	0	ND	0	0	ND	0	0.03	ND
	$\delta = 0.001$	87.5	0.73	8.3×10^{-3}	109.5	5.33	4.6×10^{-2}	122.0	13.00	9.6×10^{-2}
	$\delta = 0.25$	109.2	0.93	8.4×10^{-3}	139.6	7.49	5.1×10^{-2}	157.5	17.93	1.0×10^{-1}
	$\delta = 0.5$	149.3	1.68	1.1×10^{-2}	196.2	11.41	5.5×10^{-2}	224.9	27.52	1.1×10^{-1}
	$\delta = 0.75$	211.2	2.72	1.2×10^{-2}	290.0	20.09	6.2×10^{-2}	345.9	50.53	1.2×10^{-1}
	$\delta = 1$	295.6	0.41	1.4×10^{-3}	438.5	2.57	5.8×10^{-3}	554.5	13.07	2.3×10^{-2}

Abbreviations are as follows: TP, number of true-positive findings; FP, number of false-positive findings; rFDR, realized FDR or the proportion of false-positive SNPs among the SNPs deemed significant; Fdr, average locfdr calculated as the average locfdr of the marker set with equal or lower locfdr; ND, not defined.
[a]Number of TP and FP findings and rFDR at different XPEB locfdr significance thresholds. These are averages of 100 simulations for each genetic-architecture degree of overlap (δ); simulation parameters are the same as in Figure 2.
[b]Simulations for the null case were of 1,000 causal loci in the base GWAS and no causal loci in the target GWAS.
[c]Number of TP and FP findings and rFDR for the same simulations at different average locfdr (Fdr) thresholds.

from the base population (Figure 2D). Both meta-FE and meta-RE analyses, in contrast, are strongly driven by the larger GWAS in the base population; consequently, these methods essentially rediscover trait-associated variants in the base population (instead of discovering those in the target population) and incur an unacceptably inflated false-positive rate. At the opposite extreme, if the trait-associated variants of the target and base populations were known to completely overlap, then a meta-analysis would be the sensible procedure. However, even in such a situation, XPEB achieved an efficiency similar to that of the meta-FE and meta-RE analyses, unless a very high specificity was imposed (e.g., less than ten false positives; inset in Figure 2A). In this setting, the target-only approach suffered a substantial loss of power for ignoring the information from the base population. For traits with partially overlapping causative variants, the XPEB outperformed the target-only, meta-FE, and meta-RE approaches (Figures 2B and 2C). Simulations varying other factors, such as the sample sizes of the GWAS and the heritability of the trait, suggest that XPEB has more power than the target-only and meta-analysis approaches across a broad range of settings (Figure S2). The numbers of false positives and true positives incurred by the four methods at a conventional threshold of $p < 5 \times 10^{-8}$ or locfdr < 0.05 are compared in Table S1. Overall, XPEB, which achieves competitive or superior performance under broad scenarios without requiring a priori assumptions of the degree of overlap in genetic architecture, offers a practical strategy for analyzing a variety of traits.

Simulations with LD

The relative performance of XPEB and the target-only, meta-FE, and meta-RE approaches followed similar patterns in the presence of LD. As explained in the Material and Methods, in this simulation the decision was made for a window harboring multiple SNPs rather than a single SNP. Figure S3 and Table S2 compare power characteristics of the four methods: target only, meta-FE, meta-RE, and XPEB. As with the previous set of experiments, XPEB outperformed the other three methods when the trait-associated loci partially overlapped; it performed competitively with the best of the other three methods when trait-associated loci were either entirely overlapping or independently distributed in the target and base populations (Figure S3). We also simulated binary disease traits by using GCTA; the relative efficiency was similar when the target and base p values were derived from case-control analyses. On the basis of a previously developed argument,[31] the estimated locfdr produced by XPEB controls for the fraction of falsely rejected SNPs but not for the fraction of falsely rejected windows. However, if we retain only the minimum locfdr in each rejected window, the maximum of these minimum locfdrs remains a conservative estimate for the falsely discovered windows, and the average of these minimum locfdrs provides a good approximation of the rFDR (Table S3).

Interestingly, whereas κ_1, the parameter measuring the degree of overlap in genetic architecture, could be estimated with little bias in the absence of LD, $\widehat{\kappa}_1$ was substantially under-estimated in the presence of LD (Figure 3).

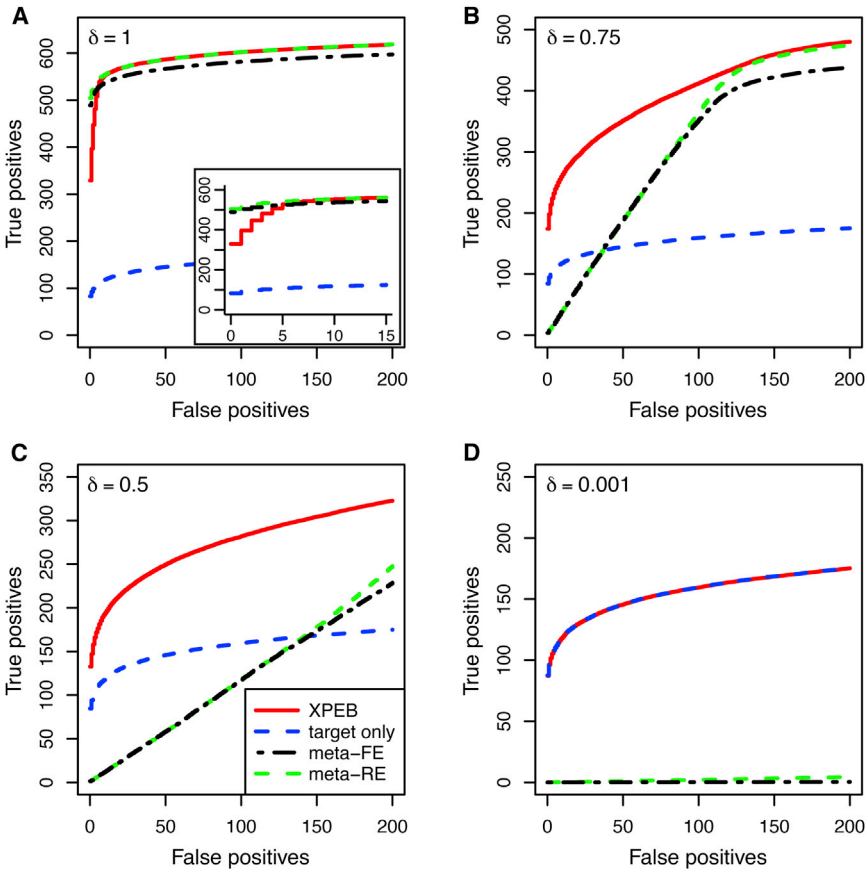

Figure 2. Partial Receiver Operating Characteristic under Independent Simulations with Varying Degrees of Overlap
Number of true-positive (y axis) and false-positive (x axis) discoveries according to four methods—XPEB, p value in the target GWAS alone (target only), p value from a fixed-effects meta-analysis of base and target GWASs (meta-FE), and p value from a random-effects meta-analysis of base and target GWASs (meta-RE produced by METASOFT)—for $\delta = 1$ (A), 0.75 (B), 0.5 (C), and 0.001 (D). Each simulated dataset consisted of 10^6 independent markers, of which 1,000 causal SNPs explained 70% of the phenotypic variance; the base GWAS had a sample size of 10^5 individuals, and the target GWAS included 10^4 individuals. Each curve is based on the average of 100 simulations. The inset in (A) indicates that meta-analysis (meta-FE and meta-RE) performed better than XPEB at very low false-positive counts (<10) when $\delta = 1$; otherwise, XPEB performed as well as or better than other methods considered.

This occurs because SNPs in strong LD with a causal variant will appear to come from the non-null components. Even if the causal variants are identical in the base and target populations, these "shoulder" SNPs might not coincide because of incongruent local LD patterns in the base and target populations. This discrepancy would be considered evidence of genetic heterogeneity, causing under-estimation of the degree of overlap and leading to a loss of power.

Plasma Lipid Concentration in AA Women

Previously, we performed genome-wide association analyses of lipid concentrations in AA and HA women from WHI-SHARe.[3] A total of 13, 15, and 5 SNPs representing 6, 5, and 3 loci, respectively, were associated with HDL, LDL, and TG, respectively. We re-analyzed this dataset by using the XPEB approach and the summary statistics from the GLGC as the base GWAS.[28] On a 2.66-GHz Intel Xeon processor, each analysis took an average of 20 min. For all three traits, XPEB estimated very high overlap in genetic architecture between EU and AA individuals: $\widehat{\kappa}_1$ achieved the maximal allowed value. At locfdr < 0.05, a total of 177, 140, and 133 SNPs at 29, 19, and 17 loci, respectively, were discovered for HDL, LDL, and TG, respectively. Without using the base GWAS, we estimated $\nu_m = P(D_m = 1|S_m)$ in a parallel fashion (as described in Equation 5), and found three, four, and two loci reaching locfdr < 0.05 for HDL, LDL, and TG, respectively. This indicates that the increased loci detected by XPEB were not due to the locfdr's being less conservative than the p-value-based genome-wide-significant criterion. The estimated FDR, computed as the average of the minimum locfdr across all significant loci, was 0.012. The 65 loci identified by XPEB included 13 of 14 loci previously identified on the basis of the p values in WHI-SHARe AA individuals alone; the single exception was a chromosome 21 HDL-associated locus ($p = 2.26 \times 10^{-8}$), which has not been replicated in another cohort[3] (Table S4 and Figure S4). Figure S5 illustrates the relationship between the p value in the WHI-SHARe AA analysis (target GWAS) and the XPEB locfdr that incorporates the association strength in the base GWAS.

We sought to validate the loci discovered by XPEB in two datasets: a cohort of 7,138 AA individuals from CARe[32] and a cohort of 3,587 HA individuals from WHI-SHARe.[3] Overall, there was strong agreement in the sign of the estimated effects: 56 of 62 loci agreed in direction between WHI-SHARe AA and CARe AA individuals (two-sided binomial test, $p = 2.97 \times 10^{-11}$), and 57 of 65 loci agreed in direction between WHI-SHARe AA and HA individuals (two-sided binomial test, $p = 3.16 \times 10^{-10}$; Table S4). Furthermore, significant loci identified by XPEB in WHI-SHARe AA individuals were enriched with small p values in WHI-SHARe HA and CARe AA individuals and showed correlated allelic effects (Figures S6 and S7). Therefore, it is reasonable to expect that a majority of the loci detected by XPEB in WHI-SHARe AA individuals will be replicated in much larger AA cohorts. Of note, two loci for each of the HDL, LDL, and TG traits did not reach $p = 5 \times 10^{-8}$ for the corresponding lipid trait in either the target GWAS[3] or the

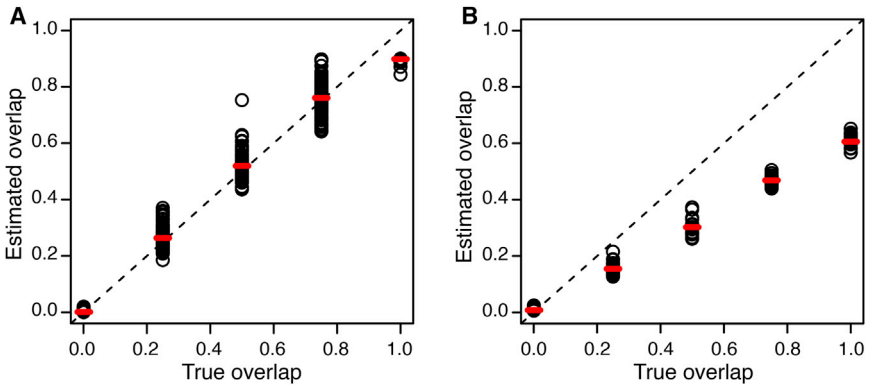

Figure 3. XPEB-Estimated Overlap Parameter Values versus the True Values
(A) The XPEB-estimated overlap parameter ($\widehat{\kappa}_1$, y axis) corresponds well to the true overlap parameters used in the simulation (δ, x axis) in simulations of independent loci. The red bar indicates the median of $\widehat{\kappa}_1$ based on 100 simulations; the simulation setting is identical to that described in the legend of Figure 2.
(B) In the presence of LD, XPEB under-estimated the degree of overlap. The red bar is the median of $\widehat{\kappa}_1$ based on 20 simulations. Each simulation included ~727,000 markers, of which 1,000 SNPs were causal and together explained 70% of the phenotypic variance. Genotypes were simulated to represent ten cohorts in the base population and one cohort in the target population (each cohort consists of 10^4 individuals). LD in the base and target populations was simulated to resemble that of HapMap CEU and YRI, respectively.

base GWAS,[28] suggesting that trans-ethnic analysis has the potential to augment our understanding of complex traits in all populations.

Cross-phenotype XPEB Analysis

One situation researchers face is that some studies use related but not identical phenotypes. For example, is it appropriate to use a study of LDL in EU populations as the base GWAS to learn about HDL in AA populations? To address this question, we applied XPEB to pairs of traits for which we had GWAS summary statistics in AA (target) and EU (base) populations. The estimated degree of overlap is shown in Table 2. For each lipid trait, when the matching trait was used as the base GWAS, the estimated κ_1 was uniformly high. When biologically related traits were analyzed together, (e.g., LDL in the base GWAS and HDL in the target GWAS), this parameter dropped sharply, suggesting pleiotropic loci but only a partially overlapping genetic basis. When seemingly unrelated traits were used, (e.g., height[29] or BMI[30] in the base GWAS and LDL in the target GWAS), the overlap parameters were essentially zero. Thus, XPEB provides a safeguard against combining incompatible phenotypes by ignoring the base GWAS.

Discussion

This research was motivated by the observation that the efficiency of genetic association studies and the accuracy of using genetic variants for predicting disease risk—whether by genotyping or by sequencing—depends critically on sample size.[33] Given the challenges in establishing well-phenotyped minority cohorts, and the empirical findings that many trait-associated loci mapped in minority populations are actually shared across ethnicities, we reasoned that a method that borrows information from populations with large cohorts could accelerate complex-trait genetic research in minority populations. However, the precise degree to which we can borrow information across populations depends on the similarity in genetic architecture between populations, which is usually unknown and is both trait and population specific.

We aimed to remove the often subjective decision of selecting and meta-analyzing compatible GWASs by developing XPEB, a principled approach that adaptively integrates results from related GWASs. XPEB is computationally efficient and uses summary statistics that are commonly available. Simulation studies suggest that when the underlying trait-associated loci partially overlap between populations, XPEB outperforms both target-only and meta-analysis approaches (Figures 2B, 2C, S3B, and S3C). When trait-associated loci are independently distributed within the base and target populations, XPEB resembles the target-only approach, and both of these methods outperform meta-analysis (Figures 2D and S3D). In cases where all trait-associated loci are shared between the target and base populations, XPEB and meta-analysis outperform the target-only approach. Although meta-analysis might achieve higher power than XPEB if the allelic effects are highly correlated, the difference is generally modest (Figures 2A and S3A).

XPEB differs from meta-analysis approaches in implementation and interpretation, and they serve different goals.[34] An implicit assumption meta-analysis makes is that there is an underlying population; the observed test statistics in a given GWAS represent a realization from a sub-population. The fixed-effects meta-analysis approach assumes a constant allelic effect among all studies, whereas the random-effects meta-analysis approach allows the allelic effects to vary around an overall population value. A recently developed trans-ethnic meta-analysis approach, MANTRA, forms clusters among multiple GWASs; subsequently, a fixed-effects approach is used to combine studies that represent genetically similar populations within a cluster, and a random-effects approach is used to combine studies across clusters.[35] We wish to emphasize that although XPEB makes use of multi-ethnic GWAS results, it is not a trans-ethnic meta-analysis approach because its goal is to identify trait-associated loci that are relevant in the target population. To define this distinction more

Table 2. Overlap and Number of Significant Loci Estimated by XPEB for the Lipid Phenotypes in the WHI-SHARe AA Cohort

	Target GWAS					
	WHI-SHARe HDL (n[a] = 7,913)		WHI-SHARe LDL (n = 7,857)		WHI-SHARe TG (n = 7,914)	
Base GWAS	Overlap[b]	Loci[c]	Overlap	Loci	Overlap	Loci
None[d]	NA	3	NA	4	NA	2
European HDL (n[e] = 99,720)	0.90	29	0.25	4	0.66	12
European LDL (n = 95,290)	0.25	7	0.90	19	0.33	3
European TG (n = 96,420)	0.68	11	0.33	5	0.90	17
European BMI (n = 123,800)	0.26	3	0.14	3	0.004	2
European height (n = 133,700)	0.10	3	0.04	3	0.12	2

NA stands for not applicable.
[a]Median sample sizes in target GWASs across markers. SNPs with fewer than 7,000 non-missing observations were removed, leaving 852,739 SNPs for all lipid phenotypes.
[b]For each specified target and base GWAS combination, the XPEB estimates of the degree of overlap (κ_1) are reported.
[c]Number of loci with locfdr < 0.05. A significant locus is defined as a set of SNPs with locfdr < 0.05 and a distance between significant SNPs < 1 Mb.
[d]Number of loci with locfdr < 0.05 on the basis of WHI-SHARe AA individuals alone.
[e]Median sample sizes in the base GWAS across markers. For BMI and height, SNPs with fewer than 100,000 non-missing observations were removed; for HDL, LDL, and TG, SNPs with fewer than 80,000 were removed. This left 621,117, 615,089, 616,513, 627,441, and 631,843 SNPs overlapping with the target GWAS SNPs for the HDL, LDL, TG, BMI, and height phenotypes, respectively.

precisely, denote the set of true trait-associated loci in the target population as $\boldsymbol{T}$ and the set of true trait-associated loci in the base population as $\boldsymbol{B}$. Fixed-effects, random-effects, and trans-ethnic meta-analysis approaches such as METASOFT and MANTRA aim to detect $\boldsymbol{T} \cup \boldsymbol{B}$. The interpretation of a significant finding is that the SNP is associated with the trait in some meta-population represented by the conglomeration of all studies (Figure S8). In contrast, XPEB focuses on a specific population designated by the target GWAS, and its goal is to identify set $\boldsymbol{T}$. Information from the base GWAS is auxiliary and is only allowed to influence the locfdr if it empirically shows compatibility with the target GWAS. Importantly, loci in set $\boldsymbol{B}$ but not in set $\boldsymbol{T}$ (base-specific traits) are considered false positives under XPEB, but they are considered true positives in trans-ethnic meta-analysis.

The practical difference between meta-analysis and XPEB can be clearly appreciated in a situation where the sample size of the base GWAS far outnumbers that of the target GWAS and the risk loci only partially overlap, a realistic situation in the analysis of a target GWAS from a minority group and a base GWAS performed in EU populations. In such a situation, meta-analysis essentially re-identifies trait-associated loci in the base population ($\boldsymbol{B}$); a fraction of these loci are not shared between populations and are therefore false positives with respect to the target population. As a result, meta-analysis incurs a high FDR even at stringent statistical significance. This is not the case for XPEB, as indicated by the decision boundary in Figure 1. A SNP association will not be declared significant without association evidence in the target population, no matter how strong the association is in the base population; however, with sufficiently strong evidence, target-specific loci ($\boldsymbol{T} \setminus \boldsymbol{B}$) can be detected without support in the base population.

Although our primary motivation was to integrate GWASs from heterogeneous populations, XPEB is also applicable when related traits might provide supportive information. Recent studies examining multiple phenotypes found evidence of pleitropic effects for groups of traits including cancers, psychiatric disorders, and metabolic syndromes.[36–38] When can GWASs from related traits be combined? Rather than making a subjective decision about the appropriateness of combining such studies, XPEB offers an objective approach to selectively borrow information from traits whose genetic basis partially overlaps that of the target trait. When we used XPEB to determine whether HDL loci are relevant for understanding the genetic architecture of LDL, we found that HDL indeed provides information about LDL but that the influence is weaker than that from an LDL study (Table 2). Therefore, we envision XPEB as broadly applicable, for example, for determining whether a GWAS of mammographic density informs risk loci for breast cancer.

The intuition of up-weighing a set of candidate SNPs on the basis of external information has been explored in the context of genetic association studies, but most studies assume either that the weighing scheme can be fixed a priori or that the set of candidate SNPs are pre-defined. For example, to weigh association results by linkage evidence, it is often reasonable to assume that the linkage peaks and association signals largely overlap; therefore, a pre-determined weighing scheme might work well.[39] Recently, several Bayesian and EB approaches have been developed for prioritizing GWAS results.[40–44] Most of these methods aim to incorporate biological information and are based on the belief that truly causal genetic variants are enriched in some pathways or share specific functional annotation. Whereas pathway and annotation information are assumed to be known without error (a gene is either in a pathway or

not), the same is not true for the association evidence in the base GWAS that XPEB tries to integrate. Whether a variant influences the trait in the base population is not directly observed in any GWAS of finite sample size; even the largest GWAS to date is underpowered to detect the many variants with small effects. Hence, one contribution of the XPEB approach is to simultaneously determine the weighing scheme and account for the imperfect information in the base GWAS. A second statistical contribution is an improved deconvolution algorithm for the mixture density in Equations 1 and 2: this algorithm achieves greater numerical stability over commonly used approaches, such as the expectation-maximization (EM) algorithm. Estimating the components of the mixture density is particularly challenging when one component dominates; in our situation, we expect that an overwhelming majority of SNPs are not associated with the trait. Consistent with the theoretical investigation,[18] our simulation indicates that the EM algorithm yields numerically unstable estimates when the non-null proportions are low. The MCMC-based algorithm we implement in XPEB ameliorates this problem by considering multiple plausible deconvolutions. This part of the algorithm is stand-alone and could be incorporated as a module in other types of mixture-deconvolution problems.

XPEB could be enhanced in several directions. The likelihood model in Equation 3 could be extended so that multiple base GWASs could be adaptively integrated simultaneously. Another area of improvement would be to better model the LD between markers. Our simulation results suggest that ignoring LD between markers leads to an underestimation of the degree of overlap and thus reduces the efficiency of the method (Figure 3). Likewise, XPEB does not explicitly model allelic heterogeneity: when a large fraction of overlapping trait-associated loci harbor population-specific trait variants, the degree of overlap can also be underestimated (data not shown). Hence, XPEB errs on the conservative side, and a model that properly accounts for LD structure and allelic heterogeneity might be able to borrow information from the base GWAS more aggressively.

Although XPEB improves our ability to map complex traits in minority populations by using a GWAS from a distantly related population with large cohorts, studies that focus on minority populations remain critical. Borrowing information from related populations cannot be a substitute for minority-specific studies. Increased sample size in the target population will improve the efficiency of XPEB in three ways. First, the power of target-specific loci can only be detected through an increased sample size in the target population. Second, for trait-associated loci that are shared between target and base populations, increased sample size in either population will improve the overall power; however, increasing the sample size in the target population is likely to bring greater marginal gain, because the base GWAS most likely outnumbers the target GWAS already. Finally, as the degree of overlap (κ_1) is estimated on the basis of the consistency in the occurrence of putative trait-associated loci in the target and base populations, increasing the sample size of the target population will improve the accuracy of the overlap parameter by bringing more true trait-associated loci into a detectable range. We also note that, as demonstrated in the lipid-trait example (Table S4), integrating GWAS results across populations by using XPEB can potentially uncover truly novel trait-associated loci that are not significant in either the target or the base GWAS. We advocate that minority-specific cohorts continue to be developed, but minority-specific GWASs should be analyzed in conjunction with related studies across ethnicities. Ultimately, by using sufficiently large sample sizes and borrowing information across ethnicities, we have the opportunity to gain a comprehensive understanding of the genetic architecture of complex traits by assembling information across populations.

Appendix A

The Null and Non-null Components of the Mixture Densities

According to the notation used in the Material and Methods, let s_j and s'_j be the observed test statistics for marker j in the target and base populations, respectively. The marginal distributions of S and S', modeled as mixture densities, are displayed as Equations 1 and 2 in the Material and Methods, respectively. In both of these mixture densities, the null distribution, g_0, is modeled as a χ^2_1 random variable, regardless of the GWAS sample size. The test statistic at a marker, m, associated with the trait follows a non-central χ^2_1 distribution with non-centrality parameters $\xi^2_m n$, where ξ^2_m measures the proportion of variance explained by the variant and n is the sample size. We assume that the marginal distributions of ξ are shared between target and base populations, but the non-centrality parameters are adjusted to account for different sample sizes.

We now describe the individual non-null components, g_j, of Equation 1. The formulation of g'_j in Equation 2 is completely analogous. We adopted a hierarchical model in which g_j has the density of a χ^2_1 random variable with a random non-centrality parameter. Let $d\text{beta}(\theta; a, b)$ denote the density at θ of a beta random variable with parameters a and b. Let $F(x;\eta)$ be the cumulative density function (CDF) at x of a χ^2_1 random variable with a non-centrality parameter of η. The CDF of g_j takes the form

$$G_j(x) = \int_0^1 F\left(x; n\frac{\exp(c\theta^2 - 1)}{\exp(c-1)}\right) d\text{beta}(\theta; j, J+1-j)d\theta.$$

In the implementation of XPEB, we let $J = 25$, $c = 9$, and n be the sample size in the target GWAS. The distribution of the non-centrality parameter, $\eta(\theta) = n\exp(c\theta^2 - 1)/\exp(c-1)$, is designed so that η ranges from 0 to n, and the mean of η increases with j. If we assume that the strongest trait-associated locus explains no

more than half of the total phenotypic variation—an assumption that holds for virtually all non-Mendelian traits—the maximum non-centrality parameter value for S at sample size n is n. Figure S1 displays a series of CDFs, G_j, ($j = 0,\cdots,25$), for $n = 100{,}000$.

Estimating Parameters in Equations 1 and 2

We next describe a MCMC algorithm for computing the posterior mean estimates of π_1, π_1', and $\boldsymbol{p}$. For this step, we treat the observed statistics, s_j and s_j', as independent draws from the mixture density in Equations 1 and 2 and seek the parameters that maximize the pseudo-likelihood function,

$$L = \prod_i g(s_i)g(s_i'). \qquad \text{(Equation A1)}$$

The priors of the parameters are specified as follows: π_1 and π_1' are assumed to be independent draws from a density

$$f_\mu(x) = (1-\gamma(\mu))x^{-\gamma(\mu)}1(0 < x \leq 1),$$

where $\gamma(\mu) = (1-2\mu)/(1-\mu)$ and $\mu = 10^{-6}$. The density function, $f_\mu(x)$, is designed to reflect a conservative belief that only a small fraction of the genome affects the trait: the density has an expectation of μ, features a spike at 0, and decays rapidly near 0. The vector $\boldsymbol{p} = (p_1,\cdots,p_J)$ is modeled as a draw from a Dirichlet of order J, Dir(α, $\cdots$, α), where α is a hyper-parameter with a uniform prior of $U(0,1)$.

The posterior distributions of the parameters are computed by MCMC via the Metropolis-Hastings algorithm.[45] Two kinds of moves are proposed: (1) random perturbation of the π_1, π_1', or α parameter and (2) perturbation of the relative non-null components, (p_{j1}, p_{j2}), subject to the constraints that $0 \leq p_j \leq 1$ for all j and $\sum p_j = 1$ before and after the move. The latter moves, which have the effects of shifting mass between non-null components, g_{j1} and g_{j2}, either choose a random pair of distinct coordinates (j_1,j_2) uniformly from 1 to J or shift mass between adjacent coordinates (j, $j+1$). Proposals that perturb π_1, π_1', or α occur with a probability of 1/15 each; mass shift between random pairs of coordinates and mass shift between adjacent coordinates are each proposed with a probability of 2/5. All simulation and data analyses presented in the Results are based on 10^6 MCMC steps. In subsequent estimation of κ_0 and κ_1, as well as the calculation of locfdr, the posterior mean of π_1, π_1', and p_j from this MCMC step was used.

Evaluating the likelihood in Equation A1 is computationally intensive because $g_j(s_m)$ does not have a closed analytic form. To reduce computational burden, we adopted a trick[46] in which the likelihood function is not evaluated with the raw s and s' values. Instead, data are binned into 500 intervals, resulting in counts c_i and c_i' for bins B_i and B_i', respectively. Let M_{ij} denote the mass assigned to bin B_i under distribution G_j, and define M_{ij}' analogously in the base GWAS. These M_{ij} and M_{ij}' are pre-computed by numerical integration and stored in two matrices. The discretized L is then $L = \prod_i (m_i)^{c_i}(m_i')^{c_i'}$, where $m_i = (1-\pi_1)M_{i0} + \pi_1\sum_{j=1}^J p_j M_{ij}$ and m_i' is defined analogously.

LD Trimming

A two-iteration parameter-estimation procedure, which includes an LD-trimming step, is introduced to correct for the biased π_1, π_1', and $\boldsymbol{p}$ estimates (due to LD) in the mixture densities. In iteration 1, the two-step algorithm described in the Material and Methods is applied to the full data without consideration of LD; next, any regions showing suggestive evidence of association, defined as $\max(\nu_m, \nu_m', 1-\omega_m) > 0.5$ for some marker m in the region, are trimmed by removal of all markers except the one that achieves the minimum ω_m in the region. In the simulation experiments, which generated haplotypes of 25-SNP windows, the underlying window boundary was used for defining regions; in the real-data analysis, SNPs showing suggestive evidence of association and within 200 kb of each other were merged into one region, given that most observed GWAS peaks span less than that range. Finally, in iteration 2, π_1, π_1', $\boldsymbol{p}$, and κ_0 are re-estimated with the trimmed dataset, and the locfdr is computed with Equation 6 in the Material and Methods for all SNPs.

Additional Explanations of Figure 1

Figure 1 displays the contours of locfdr = 0.05 for various values of κ_1 while fixing κ_0, π_1, π_1', and other simulation parameters at known values. The conventional genome-wide significance threshold of 5×10^{-8} is plotted to aid qualitative comparison, but we wish to make a conceptual distinction: although the conventional p-value-based significance criterion controls the family-wise error rate, XPEB controls the FDR;[47] the stringencies of these significance criteria are not directly comparable.[22] Furthermore, the particular threshold value of XPEB depends on the number of detectable true trait-associated loci, which in turn depends on model parameters κ_0, κ_1, π_1, and p_j. Therefore, the x intercept of XPEB decision boundaries can fall higher or lower than that of the p-value-based criterion. In Figure 1, we set $\pi_1 = \pi_1' = 0.001$, the true simulation values, and used the estimated $\widehat{\boldsymbol{p}}$ based on the WHI-SHARe LDL data.

Perhaps more curious is the observation that the XPEB decision boundaries can cross each other when κ_1 varies: the x intercept of the decision boundary for $\kappa_1 = 0.9$ is greater than the corresponding x intercept for $\kappa_1 = \kappa_0 = 0.001$. In fact, this is sensible for the following reasons: when $\kappa_1 \gg \kappa_0$, a majority of SNPs showing association evidence in the target GWAS are likely to show strong association in the base GWAS as well. Hence, the lack of evidence of association at a SNP in the base GWAS is Bayesian evidence against the hypothesis that this SNP is associated with the trait in the target GWAS. Stronger evidence of association is required in the target GWAS in order to compensate for this contrary evidence. Of course, such comparison assumes that all other model parameters are held equal.

Supplemental Data

Supplemental Data include eight figures and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.03.008.

Acknowledgments

We thank Greg Barsh for helpful discussion, two anonymous reviewers for their constructive comments, and the Genetics Bioinformatics Service Center at Stanford University for computational resources. This study was supported by US NIH grants R01 GM073059 (H.T., M.A.C., and S.I.C.), R01 HG006124 (C.K.), and R01 HG006292 (Y.L.). The Women's Health Initiative program is funded by the NHLBI (NIH, US Department of Health and Human Services) through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

Received: November 21, 2014
Accepted: March 10, 2015
Published: April 16, 2015

Web Resources

The URLs for data presented herein are as follows:

GCTA, http://www.complextraitgenomics.com/software/gcta/
HAPGEN2, http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html
METASOFT, http://genetics.cs.ucla.edu/meta/
PLINK, http://pngu.mgh.harvard.edu/purcell/plink/
XPEB, http://med.stanford.edu/tanglab/software/XPEB.html

References

1. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
2. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
3. Coram, M.A., Duan, Q., Hoffmann, T.J., Thornton, T., Knowles, J.W., Johnson, N.A., Ochs-Balcom, H.M., Donlon, T.A., Martin, L.W., Eaton, C.B., et al. (2013). Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. Am. J. Hum. Genet. *92*, 904–916.
4. Chang, M.H., Ned, R.M., Hong, Y., Yesupriya, A., Yang, Q., Liu, T., Janssens, A.C., and Dowling, N.F. (2011). Racial/ethnic variation in the association of lipid-related genetic variants with blood lipids in the US adult population. Circ Cardiovasc Genet *4*, 523–533.
5. Elbers, C.C., Guo, Y., Tragante, V., van Iperen, E.P., Lanktree, M.B., Castillo, B.A., Chen, F., Yanek, L.R., Wojczynski, M.K., Li, Y.R., et al. (2012). Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations. PLoS ONE *7*, e50198.
6. Keebler, M.E., Deo, R.C., Surti, A., Konieczkowski, D., Guiducci, C., Burtt, N., Buxbaum, S.G., Sarpong, D.F., Steffes, M.W., Wilson, J.G., et al. (2010). Fine-mapping in African Americans of 8 recently discovered genetic loci for plasma lipids: the Jackson Heart Study. Circ Cardiovasc Genet *3*, 358–364.
7. Dumitrescu, L., Carty, C.L., Taylor, K., Schumacher, F.R., Hindorff, L.A., Ambite, J.L., Anderson, G., Best, L.G., Brown-Gentry, K., Bůžková, P., et al. (2011). Genetic determinants of lipid traits in diverse populations from the population architecture using genomics and epidemiology (PAGE) study. PLoS Genet. *7*, e1002138.
8. Adeyemo, A., Bentley, A.R., Meilleur, K.G., Doumatey, A.P., Chen, G., Zhou, J., Shriner, D., Huang, H., Herbert, A., Gerry, N.P., et al. (2012). Transferability and fine mapping of genome-wide associated loci for lipids in African Americans. BMC Med. Genet. *13*, 88.
9. Carlson, C.S., Matise, T.C., North, K.E., Haiman, C.A., Fesinmeyer, M.D., Buyske, S., Schumacher, F.R., Peters, U., Franceschini, N., Ritchie, M.D., et al.; PAGE Consortium (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol. *11*, e1001661.
10. Ayala, G.X., Carnethon, M., Arredondo, E., Delamater, A.M., Perreira, K., Van Horn, L., Himes, J.H., Eckfeldt, J.H., Bangdiwala, S.I., Santisteban, D.A., and Isasi, C.R. (2014). Theoretical foundations of the Study of Latino (SOL) Youth: implications for obesity and cardiometabolic risk. Ann. Epidemiol. *24*, 36–43.
11. Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. *42*, 570–575.
12. Franceschini, N., Fox, E., Zhang, Z., Edwards, T.L., Nalls, M.A., Sung, Y.J., Tayo, B.O., Sun, Y.V., Gottesman, O., Adeyemo, A., et al.; Asian Genetic Epidemiology Network Consortium (2013). Genome-wide association analysis of blood-pressure traits in African-ancestry individuals reveals common associated genes in African and non-African populations. Am. J. Hum. Genet. *93*, 545–554.
13. Farrer, L.A., Cupples, L.A., Haines, J.L., Hyman, B., Kukull, W.A., Mayeux, R., Myers, R.H., Pericak-Vance, M.A., Risch, N., and van Duijn, C.M.; APOE and Alzheimer Disease Meta Analysis Consortium (1997). Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. JAMA *278*, 1349–1356.
14. Efron, B. (2013). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Cambridge University Press).
15. Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. The Annals of Mathematical Statistics *35*, 1–20.
16. Efron, B. (2007). Size, power and false discovery rates. Ann. Stat. *35*, 1351–1377.
17. Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. Ann. Appl. Stat. *4*, 422–438.
18. Cai, T.T., Jeng, J.X., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. J. R. Stat. Soc. Series B Stat. Methodol. *73*, 629–662.

19. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
21. Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am. J. Hum. Genet. *88*, 586–598.
22. Wang, X., Chua, H.X., Chen, P., Ong, R.T., Sim, X., Zhang, W., Takeuchi, F., Liu, X., Khor, C.C., Tay, W.T., et al. (2013). Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. Hum. Mol. Genet. *22*, 2303–2311.
23. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
24. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.
25. Consortium, T.I.H.; International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
26. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2013). Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol. Biol. *1019*, 215–236.
27. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
28. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.
29. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
30. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Mägi, R., et al.; MAGIC; Procardis Consortium (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat. Genet. *42*, 937–948.
31. Siegmund, D.O., Yakir, B., and Zhang, N.R. (2011). The false discovery rate for scan statistics. Biometrika *98*, 979–985.
32. Musunuru, K., Lettre, G., Young, T., Farlow, D.N., Pirruccello, J.P., Ejebe, K.G., Keating, B.J., Yang, Q., Chen, M.H., Lapchyk, N., et al.; NHLBI Candidate Gene Association Resource (2010). Candidate gene association resource (CARe): design, methods, and proof of concept. Circ Cardiovasc Genet *3*, 267–275.
33. Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J., and Park, J.H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. *45*, 400–405.e1–e3.
34. Evangelou, E., and Ioannidis, J.P. (2013). Meta-analysis methods for genome-wide association studies and beyond. Nat. Rev. Genet. *14*, 379–389.
35. Morris, A.P. (2011). Transethnic meta-analysis of genomewide association studies. Genet. Epidemiol. *35*, 809–822.
36. Setiawan, V.W., Schumacher, F., Prescott, J., Haessler, J., Malinowski, J., Wentzensen, N., Yang, H., Chanock, S., Brinton, L., Hartge, P., et al. (2014). Cross-cancer pleiotropic analysis of endometrial cancer: PAGE and E2C2 consortia. Carcinogenesis *35*, 2068–2073.
37. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
38. Carty, C.L., Bhattacharjee, S., Haessler, J., Cheng, I., Hindorff, L.A., Aroda, V., Carlson, C.S., Hsu, C.N., Wilkens, L., Liu, S., et al. (2014). Analysis of metabolic syndrome components in >15 000 african americans identifies pleiotropic variants: results from the population architecture using genomics and epidemiology study. Circ Cardiovasc Genet *7*, 505–513.
39. Roeder, K., Bacanu, S.A., Wasserman, L., and Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. Am. J. Hum. Genet. *78*, 243–252.
40. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
41. Chen, G.K., and Witte, J.S. (2007). Enriching the analysis of genomewide association studies with hierarchical modeling. Am. J. Hum. Genet. *81*, 397–404.
42. Carbonetto, P., and Stephens, M. (2013). Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. PLoS Genet. *9*, e1003770.
43. Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am. J. Hum. Genet. *86*, 6–22.
44. Zablocki, R.W., Schork, A.J., Levine, R.A., Andreassen, O.A., Dale, A.M., and Thompson, W.K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. Bioinformatics *30*, 2098–2104.
45. Liu, J.S. (2008). Monte Carlo Strategies in Scientific Computing (Springer).
46. Efron, B. (2008). Microarrays, empirical bayes and the two-groups model. Stat. Sci. *23*, 1–22.
47. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.